DOCUMENT RESUME

ED 402 558 CS 012 669

AUTHOR Valencia, Sheila W.; Au, Kathryn H.
TITLE Portfolios across Educational Contexts: Issues of Evaluation, Teacher Development, and System Validity. Reading Research Report No. 73.
INSTITUTION National Reading Research Center, Athens, GA.; National Reading Research Center, College Park, MD.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 97
CONTRACT 117A20007
NOTE 42p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Case Studies; Elementary Education; *Literacy; *Portfolio Assessment; Portfolios (Background Materials); Professional Development; Program Effectiveness; *Student Evaluation; Validity
IDENTIFIERS Bellevue School District WA; Kamehameha Early Education Program

ABSTRACT

A case study across 2 different elementary education settings examined (1) how well portfolios document literacy learning that is both authentic and aligned with curriculum; (2) teachers' ability to interpret and evaluate portfolio evidence from more than one site; and (3) what teachers learn about literacy instruction and assessment as a result of cross-site collaboration. The two programs were the Bellevue Literacy Portfolio Project (located in a suburb of Seattle, Washington) and the Kamehameha Elementary Education Program (a privately funded educational research and development effort in Hawaii). Results suggest that portfolios contained authentic artifacts of students' literacy experiences, although there was a substantial amount of evidence judged to be missing from the portfolios. Nevertheless, with a shared understanding of literacy learning, teachers were able to reach a high degree of agreement when rating portfolios from different sites and enhance their understanding of both learning and assessment through the cross-site evaluation process. Findings should not be interpreted simply as findings on portfolio assessment--they must be interpreted in light of a complete portfolio system in which attention is given to generating and collecting artifacts, supporting collaborative evaluation, and providing ongoing professional development. Supportive internal and external conditions must be present if portfolios are to become effective tools for literacy assessment and professional development. (Contains 48 references, and 5 tables and 3 figures of data.) (Author/RS)

Portfolios Across Educational Contexts: Issues of Evaluation, Teacher Development, and System Validity

Sheila W. Valencia
University of Washington

Kathryn H. Au
University of Hawaii

NRRC

National
Reading Research
Center

READING RESEARCH REPORT NO. 73
Winter 1997

NRRC

National Reading Research Center

Portfolios Across Educational Contexts: Issues of Evaluation, Teacher Development, and System Validity

Sheila W. Valencia
University of Washington

Kathryn H. Au
University of Hawaii

READING RESEARCH REPORT NO. 73
Winter 1997

The work reported herein is a National Reading Research Center Project of the University of Georgia and University of Maryland. It was supported under the Educational Research and Development Centers Program (PR/AWARD NO. 117A20007) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed here do not necessarily reflect the position or policies of the National Reading Research Center, the Office of Educational Research and Improvement, or the U.S. Department of Education.

NRRC National Reading Research Center

NRRC - University of Georgia
318 Aderhold
University of Georgia
Athens, Georgia 30602-7125
(706) 542-3674 Fax: (706) 542-3678
INTERNET: NRRC@uga.cc.uga.edu

NRRC - University of Maryland College Park
3216 J. M. Patterson Building
University of Maryland
College Park, Maryland 20742
(301) 405-8035 Fax: (301) 314-9625
INTERNET: NRRC@umail.umd.edu

About the National Reading Research Center

The National Reading Research Center (NRRC) is funded by the Office of Educational Research and Improvement of the U.S. Department of Education to conduct research on reading and reading instruction. The NRRC is operated by a consortium of the University of Georgia and the University of Maryland College Park in collaboration with researchers at several institutions nationwide.

The NRRC's mission is to discover and document those conditions in homes, schools, and communities that encourage children to become skilled, enthusiastic, lifelong readers. NRRC researchers are committed to advancing the development of instructional programs sensitive to the cognitive, sociocultural, and motivational factors that affect children's success in reading. NRRC researchers from a variety of disciplines conduct studies with teachers and students from widely diverse cultural and socioeconomic backgrounds in pre-kindergarten through grade 12 classrooms. Research projects deal with the influence of family and family-school interactions on the development of literacy; the interaction of sociocultural factors and motivation to read; the impact of literature-based reading programs on reading achievement; the effects of reading strategies instruction on comprehension and critical thinking in literature, science, and history; the influence of innovative group participation structures on motivation and learning; the potential of computer technology to enhance literacy; and the development of methods and standards for alternative literacy assessments.

The NRRC is further committed to the participation of teachers as full partners in its research. A better understanding of how teachers view the development of literacy, how they use knowledge from research, and how they approach change in the classroom is crucial to improving instruction. To further this understanding, the NRRC conducts school-based research in which teachers explore their own philosophical and pedagogical orientations and trace their professional growth.

Dissemination is an important feature of NRRC activities. Information on NRRC research appears in several formats. *Research Reports* communicate the results of original research or synthesize the findings of several lines of inquiry. They are written primarily for researchers studying various areas of reading and reading instruction. The *Perspective Series* presents a wide range of publications, from calls for research and commentary on research and practice to first-person accounts of experiences in schools. *Instructional Resources* include curriculum materials, instructional guides, and materials for professional growth, designed primarily for teachers.

For more information about the NRRC's research projects and other activities, or to have your name added to the mailing list, please contact:

Donna E. Alvermann, Co-Director
National Reading Research Center
318 Aderhold Hall
University of Georgia
Athens, GA 30602-7125
(706) 542-3674

John T. Guthrie, Co-Director
National Reading Research Center
3216 J. M. Patterson Building
University of Maryland
College Park, MD 20742
(301) 405-8035

NRRC Editorial Review Board

Susan Lytle
University of Pennsylvania

Bert Mangino
Las Vegas, Nevada

Susan Mazzoni
Baltimore, Maryland

Ann Dacey McCann
University of Maryland College Park

Sarah McCarthey
University of Texas at Austin

Veda McClain
University of Georgia

Lisa McFalls
University of Georgia

Randy McGinnis
University of Maryland

Mike McKenna
Georgia Southern University

Barbara Michalove
Fourth Street Elementary School
Athens, Georgia

Elizabeth B. Moje
University of Utah

Lesley Morrow
Rutgers University

Bruce Murray
Auburn University

Susan Neuman
Temple University

John O'Flahavan
University of Maryland College Park

Marilyn Ohlhausen-McKinney
University of Nevada

Penny Oldfather
University of Georgia

Barbara M. Palmer
Mount Saint Mary's College

Stephen Phelps
Buffalo State College

Mike Pickle
Georgia Southern University

Amber T. Prince
Berry College

Gaoyin Qian
Lehman College-CUNY

Tom Reeves
University of Georgia

Lenore Ringler
New York University

Mary Roe
University of Delaware

Nadeen T. Ruiz
California State University-
Sacramento

Olivia Saracho
University of Maryland College Park

Paula Schwanenflugel
University of Georgia

Robert Serpell
University of Maryland Baltimore
County

Betty Shockley-Bisplinghoff
Barnett Shoals Elementary School
Athens, Georgia

Wayne H. Slater
University of Maryland College Park

Margaret Smith
Las Vegas, Nevada

Susan Sonnenschein
University of Maryland Baltimore
County

Bernard Spodek
University of Illinois

Bettie St. Pierre
University of Georgia

Steve Stahl
University of Georgia

Roger Stewart
Boise State University

Anne P. Sweet
Office of Educational Research
and Improvement

Louise Tomlinson
University of Georgia

Bruce VanSledright
University of Maryland College Park

Barbara Walker
Eastern Montana University-Billings

Louise Waynant
Prince George's County Schools
Upper Marlboro, Maryland

Dera Weaver
Athens Academy
Athens, Georgia

Jane West
Agnes Scott College

Renee Weisburg
Elkins Park, Pennsylvania

Allan Wigfield
University of Maryland College Park

Shelley Wong
University of Maryland College Park

Josephine Peyton Young
University of Georgia

Hallie Yopp
California State University

About the Authors

Sheila Valencia is Associate Professor of Curriculum and Instruction at the University of Washington, Seattle. She received her doctorate in reading education from the University of Colorado, Boulder, and then returned to the public schools for 6 years as a district reading specialist. Her research focuses on literacy instruction and assessment, with a special emphasis on classroom-based assessment. Dr. Valencia has served on the advisory boards of several national, state, and professional assessment projects including the National Academy of Education, National Task Force on Assessment, and the Joint Committee on Assessment of NCTE and IRA.

Kathryn H. Au is Associate Professor in the College of Education at the University of Hawaii at Manoa. She received her doctorate in educational psychology from the University of Illinois at Urbana-Champaign, after working as a classroom teacher in the primary grades. Her research focuses on the literacy instruction of students of diverse cultural and linguistic backgrounds. She is president elect of the National Reading Conference and was elected a vice president of the American Educational Research Association.

National Reading Research Center
Universities of Georgia and Maryland
Reading Research Report No. 73
Winter 1997

Portfolios Across Educational Contexts: Issues of Evaluation, Teacher Development, and System Validity

Sheila W. Valencia
University of Washington

Kathryn H. Au
University of Hawaii

Abstract. *We report here a case study of literacy portfolios across two different settings. Specifically, we investigated (1) how well portfolios document literacy learning that is both authentic and aligned with curriculum; (2) teachers' ability to interpret and evaluate portfolio evidence from more than one site; and (3) what teachers learn about literacy instruction and assessment as a result of cross-site collaboration. Results suggest that portfolios contained authentic artifacts of students' literacy experiences, although there was a substantial amount of evidence judged to be missing from the portfolios. Nevertheless, with a shared understanding of literacy learning, teachers were able to reach a high degree of agreement when rating portfolios from different sites and enhance their understanding of both learning and assessment through the cross-site evaluation process. We suggest that the results should not be interpreted simply as findings on portfolio assessment. They must be interpreted in light of a complete portfolio system in which attention is given to generating and collecting artifacts, supporting collaborative evaluation, and providing ongoing professional development. Supportive internal and external conditions must be present if portfolios are to become effective tools for literacy assessment and professional development.*

All across the United States, interest in portfolios is running high. Private foundations, states, and school districts are investing millions of dollars in portfolio projects they hope will enhance teaching, learning, and assessment practices (Pelavin, 1991). The need to establish ongoing, classroom-based assessment has become a predictable part of the assessment conversation, and portfolios appear to be a promising candidate for the job.

There are three major expectations for portfolios. First, portfolios are viewed as being more meaningful, authentic, and valid indicators of what students know and can do than more traditional assessments. They can be integrated with classroom instruction, reflect the rich and complex work that children actually do, and address broader, more important learning outcomes (Au, Scheu, Kawakami, & Herman, 1990; Calfee & Perfumo, 1996;

Valencia, 1990; Wiggins, 1989a; Wolf, 1989). By including multiple indicators of student performance, portfolios also capture the variability and patterns, across tasks and time, that characterize true learning (Messick, 1994; Valencia, 1990; Wiggins, 1993).

Second, as a result of these authentic, classroom-based features, portfolios have the potential to enhance both teaching and learning. Because they are housed in the classroom and can be used regularly by teachers and students as part of the instructional program, portfolios have potential to provide more useful, meaningful, and accessible information than typical assessments. Teachers and students should become more reflective and knowledgeable as a consequence of using portfolios (Arter & Spandel, 1992; Darling-Hammond & Ancess, 1993; Johnston, 1989; Moss et al., 1992; Valencia & Calfee, 1991; Wiggins, 1989b; Wolf, 1989).

Third, some educators and psychometricians are hopeful that portfolios will provide useful assessment information for reporting to people outside the classroom or local context. This will require an acceptable level of interrater agreement as portfolio raters outside the local site examine and score portfolios. For some, this outside reporting is necessary to ensure the survival of portfolios as an assessment innovation (Calfee & Perfumo, 1996; Freedman, 1993; Valencia, 1991); for others, it is a way to enter classroom information into the policy arena (Chittenden & Spicer, 1993; LeMahieu, Eresh, & Wallace, 1992; Wixson, Valencia, & Lipson, 1994); and for others, the evaluation process itself is valued as a powerful mechanism for professional development (Wolf, LeMahieu, & Eresh, 1992).

Such an ambitious agenda for portfolio assessment poses enormous challenges for educators. It is a relatively simple task to collect evidence of student work, a much more complex one to assure that work represents authentic instances of learning, feeds back to improve teaching and learning, and yields useful assessment information (Aschbacher, 1994). The complexity is magnified and the task becomes more daunting as portfolio contents and processes vary, and participants represent a range of schools, districts, and educational contexts. As educators make site-based decisions to collect different types of portfolio artifacts, use different evaluation schemes, and participate in different levels of professional development, it may become more difficult to use portfolio data to analyze student achievement and to improve teaching and learning.

We report here a case study of literacy portfolios across two settings. Specifically, we explore: (1) the potential of literacy portfolios from different sites to capture curriculum outcomes that are both authentic and aligned with instruction; (2) teachers' ability to interpret and evaluate portfolio evidence from more than one site; and (3) what teachers learn about literacy instruction and assessment through the process of cross-site collaboration. By examining these issues, we hope to shed light on a conceptual framework for a portfolio system—the components and the external and internal conditions that are needed for portfolios to be effectively implemented and used. Specifically, we hypothesize that if the collec-

tion of portfolio artifacts, evaluation process, and professional development all receive concurrent support, and are situated in a shared understanding of literacy and a low-stakes environment, classroom-based portfolio assessment can be successfully implemented, used, and evaluated across sites. Furthermore, we hypothesize that the process of evaluating portfolios across sites provides teachers with an important "outside" perspective on instruction and assessment that transfers to their own classrooms.

Background

It is difficult to generalize about the role of portfolios in education because portfolios are defined and implemented differently within projects and sites (Calfee & Perfumo, 1996; Valencia & Calfee, 1991). In some settings, portfolios are conceptualized simply as collections of students' work. These collections provide evidence of the work students have done in the classroom, but are not the object of reflection and systematic analysis by students or teachers. In other settings, where portfolios have evolved beyond collections, portfolios serve a variety of purposes. For example, showcase portfolios are used to highlight the best work that students have done, documentation portfolios may be used to show students' growth over time with respect to specific outcomes, and evaluation portfolios may contain prespecified pieces that are each evaluated. In some settings, students' ownership of portfolios is considered of primary importance, and students, rather than teachers, decide what will go into the portfolios (Hansen, 1994; Howard, 1990; Tierney, Carter, & Desai, 1991). In other settings, portfolios are used for large-scale evaluation; teachers have guidelines about the kinds of work that should be present in each student's portfolio (Chittenden & Spicer, 1993; Koretz, Stecher, Klein, & McCaffrey, 1994; LeMahieu et al., 1992). And in others, decisions about portfolio contents are shared by the teacher and the student (Valencia & Place, 1994a). These issues of purpose and audience present important choices and implications for student and teacher ownership, portfolio contents, and use of portfolio information (Moss et al., 1992; Valencia & Calfee, 1991).

A similar range of possibilities is found in the various contexts in which portfolios are implemented. Some portfolios are simply implemented by individual teachers interested in portfolios; others are part of district or state-wide assessment systems. Some have no stakes or sanctions attached to their implementation or results while others have high public visibility and consequences. There is also great variability in the support provided to teachers for implementing and using portfolios, from minimal professional development to long-term support over years.

There is considerably less information about issues of evaluation and reporting of portfolio information than there is about implementation. Indeed, some would reject any type of evaluation or reporting, finding it undesirable, unnecessary, and potentially detrimental to the portfolio philosophy (Carini, 1975; Johnston, 1989). Others approach portfolio evaluation with interest in collaborative interpretation and narrative reporting of information. This approach cautions against relying too heavily on

"traditional" interrater agreement. It emphasizes the value of critical dialogue about collections of student work among those most knowledgeable about the context in which the assessment occurs (Johnston, 1989; Moss, 1994; Moss et al., 1992). In contrast, a few have tried to apply evaluation rubrics to portfolios, concerning themselves with issues of interrater agreement and large-scale reporting (Chittenden & Spicer, 1993; Gearhart, Herman, Baker, & Whittaker, 1992, 1993; Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; LeMahieu et al., 1992; Nystrand, Cohen, & Martinez, 1993). In general, they find questionable levels of interrater agreement and low correlations with on-demand tasks. Furthermore, interrater agreement may drop below acceptable levels for individual and program decisions unless individual items in the portfolio are scored one at a time and items are standardized (Moss, 1994). Alternatively, some evaluation and reporting has been attempted on a smaller scale, with an eye toward developing responsible evaluation and professional development rather than high stakes reporting (Au, 1994; Valenicia & Place, 1994a, 1994b). Psychometricians caution that to provide information that will satisfy those historically interested in standardized test scores will require a reconceptualization of the critical attributes of good assessment, including reliability and validity, and consideration of the different needs of various stakeholders (Haney, 1991; Linn, Baker, & Dunbar, 1991; Moss et al., 1992).

In short, there is presently no common philosophy underlying the use of portfolios, no one blueprint for how portfolios can or should be implemented in classrooms, and no one approach to professional development or to evaluation. As educators begin to explore the potential of portfolios, these various models, purposes, and contexts will be critical variables to consider. Unlike assessments of the past or even the newer on-demand performance assessments, the variability in portfolio assessment may introduce new challenges for their implementation, use, and aggregation of information within a site and, most especially, across different sites. If, however, the variability is valued and grounded in teachers' experience, knowledge, and professional development, we may be able to create portfolio systems that support teaching, learning, and assessment.

Method

The Context

The two programs participating in the study were the Bellevue Literacy Portfolio Project and the Kamehameha Elementary Education Program (KEEP). Similarities and difference between the sites are summarized in Table 1. Bellevue is a suburb of Seattle. The school district serves approximately 15,000 students, 20% of whom are minority. As a whole, test scores reflect achievement at about the 60th percentile in reading and writing. The district has a history of strong, on-going professional development. KEEP was a privately funded educational research and development effort. At the time of this study, the program served 5,300 Native Hawaiian students in 9 public schools in low-income communities on 3 of the Hawaiian islands. The test scores for students in the KEEP program were typically below the

Table 1
Similarities and Differences Across Sites

	KEEP	Bellevue
Student population	5,300	15,000
Diversity	100% minority	20% minority
Achievement	lowest 25%ile	60%ile
Literacy philosophy	Constructivist—literature-based; process writing	Constructivist—literature-based; process writing
Literacy outcomes	Reading Ownership (including variety) Reading Process (understanding, response, strategies) Writing Ownership Writing Process (quality, variety, process)	Reading Ownership Reading Variety Reading Ability (understanding, response) Reading Strategies Writing Ownership Writing Ability (quality, variety) Writing Process Self-Reflection/Evaluation
Literacy & portfolio professional development activities	at least 4 years	4 years
Teachers' portfolio implementation	at least 2 years	at least 2 years
Portfolio development	Designed by KEEP curriculum developers and consultants	Designed by teachers and district specialist
Portfolio requirements	Emphasis on required pieces, some teacher choice	Emphasis on teacher choice and student choice, some required pieces
Portfolio use	Used primarily for evaluation, administrative reports; some individual teacher use	Used primarily by classroom teachers with students; selected scoring for administrative reporting
Portfolio evaluation criteria	Lists of specific benchmarks for each of the 4 outcomes	Descriptive analytic rubric aligned with 8 outcomes
Evaluation experience	Some experience with scoring	Some experience with scoring
Portfolio participation	Voluntary	Voluntary

25th percentile. Beyond these demographics, the two sites for this project have important similarities and differences that make them a best-case scenario for a cross-site study of portfolio assessment.

Similarities. Both portfolio projects had been in place for several years and involved only teachers who had volunteered. At the time of this project, the Bellevue teachers had already worked with portfolios for 3 years—experimenting for 1 year and then implementing for 2 years. The project began with 32 volunteer teachers; by the third year, an additional 240 teachers also had participated in portfolio inservice activities. The 150 KEEP teachers had been introduced to portfolios in 1989 and had implemented portfolios in their classrooms for 2 years at the time of the project.

Project teachers at both sites were supported through on-going professional development that enabled them to learn about and share their work with portfolios. In Bellevue, a core group of approximately 30 teachers met almost monthly over 3 years. The school district charged this group with developing a model for portfolios. In addition, these teachers played a key role in disseminating portfolio information to their building colleagues. At KEEP, consultants based at each school provided teachers with knowledge about portfolios. During the first 3 years of implementation, consultants assumed major responsibility for implementing portfolios. Teachers took over responsibility for the portfolios after the third year.

Both portfolio projects grew out of efforts to change the overall language arts curriculum. These efforts resulted in new outcomes for student learning. In Bellevue, the district adopted new language arts student learning objectives (SLOs) in the fall of 1990. They were organized by age bands and broad literacy outcomes. KEEP previously had a comprehension-oriented curriculum, but changed in 1989 to a whole literacy curriculum with students' ownership of literacy as the overarching goal (Au et al., 1990). In both sites, the new curricula reflected a process-oriented, constructivist philosophy about literacy and literacy learning (Applebee, 1991; Weaver, 1990). In this philosophy, the affective and cognitive dimensions of literacy are considered to be equally important, and ownership of literacy is seen as the overarching goal. Students engage in authentic literacy activities, such as the reading of literature and writing for different purposes, that have meaning outside the classroom as well as within it. Literacy learning proceeds as a social process in which students learn from peers as well as from the teacher. Both the Bellevue and KEEP curricula included outcomes for student learning in ownership, reading, and writing. Teachers at both sites taught writing following a process approach (Calkins, 1994; Graves, 1983), engaging students in planning, drafting, revising, editing, and publishing. Literature-based instruction was the approach for teaching reading, with an emphasis on discussions of literature and writing in response to literature (Roser & Martinez, 1995; Short & Pierce, 1990).

In both sites, curriculum changes were accompanied by a desire to change the assessment system. Both sites had a commitment to large-scale evaluation as well as to individual

student progress. In both sites, portfolios were seen as a potential tool for evaluation and for professional development. At the time portfolio assessment was introduced at Bellevue and KEEP, it was not known whether one system could in fact meet both purposes. However, the portfolio systems evolved with the intention of trying to do so. Each system had specific required portfolio artifacts and an evaluation system designed to evaluate progress toward the outcomes.

Differences. The sites had important differences as well. In addition to serving different student populations, the portfolio models were different. Each was designed to meet the unique needs of the site. The Bellevue portfolio was designed to serve three functions—to evaluate student progress toward each of eight literacy outcomes, to involve students in self-reflection and self-evaluation, and to document student growth over time (Valencia & Place, 1994a). Although Bellevue wanted to see if their portfolios could be used to report district-wide student achievement, the primary emphasis was on improving instruction and engaging students in learning at the classroom level. In contrast, the KEEP portfolio was designed primarily to meet the requirements of large-scale assessment. Students were judged to be at, above, or below grade level on benchmarks for each of six literacy outcomes (Au, 1994). Consequently, portfolios from Bellevue and KEEP required different types of evidence and different levels of student and teacher involvement.

Another important difference was the manner in which portfolios were introduced into each system. The Bellevue portfolio project used a bottom-up approach. The Bellevue system was designed by a group of teachers working with the District Language Arts Specialist and a consultant. Both the School Board and the Superintendent supported the project with teacher released days each month, and they gave a 5-year commitment to the project. The group worked on the system for 1 year, revising and refining it, before they agreed to implement it, although still on a voluntary basis. The actual portfolios included some required "common tools" but varied considerably across classrooms since individual teachers and students had a major role in the selection of portfolio artifacts. A small group of teachers participated in the development, training, and implementation of a portfolio evaluation system during the third and fourth years of the project. Only a subset of portfolios was scored each year, and the results were not formally reported or used for district evaluation. The voluntary nature of the project and the low stakes associated with it, resulted in a wide of range of teacher compliance. Some teachers experimented with portfolios throughout the project, implementing them as they desired. Others, especially those involved in the evaluation, were much more rigorous and tried to adhere to the portfolio requirements (Valencia & Place, 1994a).

In contrast, KEEP portfolio implementation was a top-down process. The KEEP portfolio system was developed as part of an overhaul of the instructional program triggered by inconsistent standardized achievement test results. Although KEEP administrators and consultants were under pressure to bring change about quickly, teachers' participation in portfolios

was gradual and supported (Au, 1994). The portfolio assessment system was designed by KEEP curriculum developers and consultants, and classroom teachers were asked to implement the system. These consultants worked with the teachers on a weekly basis on issues of classroom organization, literacy instruction, and portfolio assessment. Initially, consultants did most of the collecting of portfolio information during their classroom visits; because of this consultant role and because of the nature of the KEEP learner benchmarks, observation checklists were a cornerstone of the portfolios. Consultants visited the classrooms and, with the help of the teacher, identified evidence of student performance on specific benchmarks and recorded it on the checklists. Some of the evidence came from student work, other evidence came from the consultants' classroom observations. Teachers and students had minimal engagement with the portfolios early on. In the fourth year of the project, however, teachers assumed more responsibility for keeping and selecting work in the portfolios, although there were still requirements for certain types of evidence (Au & Asam, in press).

This study investigated what appeared to be a best-case scenario for exploring portfolios across sites, in terms of both the similarities and differences in the projects. Both projects grew from similar research-based, constructivist philosophies on literacy learning and teaching, and teachers at each site knew how to create portfolios that provided rich representations of their students' literacy learning that aligned with their curricula. Teachers at both sites had considerable experience with portfolios and had moved beyond concern with the logistics of portfolio implementation. In addition, both sites had portfolio models that, from their inception, included an evaluation and reporting component. This interest in evaluation grew from concerns at the sites themselves, rather than being imposed by an external agency, such as a state or federal department of education. Therefore, no site was asked to believe or do anything that was not consistent with its original mission. This situation maximized the potential for capturing authentic indicators of students' reading and writing and for reliable evaluation across the two sites, despite differences in portfolio models.

This project was also considered to be a best-case scenario in terms of the opportunities available for studying professional development through portfolio evaluation. Due to the similarities in language arts philosophy, it was expected that teachers from the two projects would have shared understandings that would give them a common language for talking about portfolios, and the ability to understand and benefit from the contents of each others' portfolios. They would most likely have common concerns and be able to engage in meaningful, useful professional discussions. In addition, because both portfolio systems already included an evaluation component, teachers were accustomed to and comfortable with making judgments about student work.

Finally, the differences in geographic location, student population, and actual portfolio models make this case study a comparison of two like-minded but not identical systems. We expected and found considerable variability in the types of information in the portfolios.

In sum, the common literacy philosophy, years of experience implementing portfolios, and orientation to both classroom and outside reporting combined with the difference in actual portfolio requirements and student populations made this pair of sites the best case to see if artifacts across sites were meaningful, and if cross-site evaluation was desirable and feasible. These two sites also provided an excellent opportunity to look at the professional growth of experienced portfolio teachers as they learned about how portfolios worked in another context.

Participants

Four teachers (2 primary, 2 intermediate) from each site participated in this in-depth study. The 4 Bellevue teachers had from 5 to 29 years of teaching experience; 3 of the teachers had more than 13 years. All had been members of the core portfolio team and had participated in several professional development classes, workshops, and district committees. They credited these experiences and their teaching colleagues with their own professional growth. The KEEP teachers had from 21 to 28 years of teaching experience. They had been associated with KEEP for 4 to 16 years. As part of their involvement with KEEP, all the teachers had received individual feedback on their teaching from KEEP staff members, attended numerous workshops, engaged in professional reading, observed in other teachers' classrooms, and participated in teacher networks. All had implemented the whole literacy curriculum at a high level. Although all taught both a writers' and a readers' workshop, 3 of the teachers had focused their portfolio data collection on writing, while 1 had focused on reading. All 8 teachers were selected because their classrooms reflected a good understanding of their local literacy curriculum and portfolios. All had worked with portfolio assessment for at least 2 years.

Procedures

At the beginning of the study, the teachers were interviewed about their teaching backgrounds, professional development, language arts instruction, and portfolios. Teachers were interviewed again at the end of the study to determine what they had learned about portfolios and how their participation in the project had influenced their professional development. There were two cross-site meetings, one in January and one in April. The procedure for these meetings was similar (see Table 2). After getting acquainted, the visiting teachers observed in the home-site teachers' classrooms for a morning. The purpose of the observation was to orient the visitors to the particular cultural context of the classroom and to allow them to become acquainted with instructional practices reflected in the portfolios they were going to review. Following the observation, the visiting teachers met as a group with the local site director to discuss their observations.

The following 2 to 3 days, teachers evaluated portfolios from both sites using specific evaluation rubrics. At the first cross-site meeting, held at KEEP, teachers evaluated Bellevue and KEEP portfolios using the KEEP evaluation system. At the second meeting, held in Bellevue, they used the Bellevue rubric. In

Table 2
Procedure for Cross-Site Collaboration and Portfolio Evaluation

Pre-meeting interviews

↓

Meeting at KEEP
- classroom observations & discussions
- orientation to portfolios from both sites (collaborative descriptive discussions)
- training using KEEP evaluation criteria
- evaluation of KEEP and Bellevue portfolios using KEEP evaluation criteria
- collaborative cross-site discussions

↓

Meeting at Bellevue
- classroom observations & discussions
- training using Bellevue evaluation system
- evaluation of KEEP and Bellevue portfolios using Bellevue evaluation criteria
- teacher development of Common evaluation criteria
- evaluation of KEEP and Bellevue portfolios using Common evaluation criteria
- collaborative cross-site discussions

↓

Post-meeting interviews

addition, they collaboratively developed a new set of Common criteria drawing from both sets of learning outcomes and evaluation systems, and from their experiences working together. They applied the Common criteria to portfolios from both sites. Eight portfolios were evaluated using each evaluation criteria. Finally, at the close of each meeting, there was a discussion and debriefing about what the teachers had learned about portfolios and gained from the experience of working with teachers from another site.

For the purposes of portfolio evaluation, teachers worked in cross-site groups of 4—the 4 primary teachers comprised one group and the 4 intermediate teachers comprised the other. At the first cross-site meeting, teachers spent 2 hr familiarizing themselves with the portfolios from each site. Each teacher offered one of her portfolios for others to review. Using a descriptive process in which collaborative dialogue is featured (Moss, 1994; Valencia & Place, 1994b), each teacher observed as the others discussed their interpretations of her student's portfolio. The teacher entered the discussion at the end, offering her confirmations, additions, and disagreements to the conversation. After portfolios from both sites

had been discussed, teachers synthesized the ways in which the portfolio evidence from the two sites was similar and different. Often, work was linked to what the teachers had observed in the classrooms the previous day. They discussed why certain pieces were included in the portfolios and the changes they might want to make in their own portfolios. This orientation occurred only at the first meeting. It was a critical step for clarifying the evidence and the learning outcomes. By the second meeting, teachers were very familiar with the portfolio models and evidence from both sites.

Training for the actual evaluation began with an overview of the evaluation system to be used, followed by practice rating one portfolio from each site. With the initial descriptive discussions as background, differences in interpretation of evidence and performance criteria were discussed and negotiated. Formal evaluation occurred next. Each teacher contributed one portfolio to be evaluated by the 4 members of her cross-site group, making a total of four portfolios to be evaluated by each group. Teachers evaluated each portfolio independently, spending approximately 20 to 30 min with each. They read through all the pieces in the portfolio, looking within and across pieces for evidence of students' abilities on each of the learning outcomes or benchmarks. Then they recorded a rating for each. These ratings were used in the analyses. After rating all the portfolios on her own, each teacher discussed her ratings with her same site partner and then with the cross-site group of 4.

Data Sources

Portfolio evidence. The evaluated portfolios were reviewed at each site meeting and a table of contents compiled for each. The contents were analyzed to determine variability across sites, alignment with outcomes and portfolio requirements, and changes in portfolio contents over time. Pieces were coded according to the major disciplinary *focus* of each artifact: reading, writing, and other subjects (e.g., science, social studies). For example, reading response journals, summaries, book reports, and teachers' notes on students' oral reading (i.e., running records) were coded as reading artifacts. Writing artifacts included different types of writing (i.e., stories, poems, reports, journals), planning notes, rough drafts, final copies, and samples of daily oral language activities. Pieces were also coded according to the *type* of artifact: student work (e.g., reports, interest surveys, writing); anecdotal notes/checklists (e.g., teacher records, running records); other (e.g., photos, audiotapes, artwork); student-selected pieces/goals/self-evaluation (e.g., student entry slips on portfolio items, self-evaluation forms, personal goals); and parent input (questionnaires, comments). Some artifacts were double coded. For example, a piece of writing with rough draft and a student entry slip describing why the student selected the piece for the portfolio was coded as writing and student-selected/self-reflection; a final draft of a book report was coded as reading and writing; a research report on salmon was coded as writing and as other content area; and running records were coded as reading and as anecdotal notes in order to reflect both the focus of the activity and the documentation mode. Most of the reading artifacts were written responses (i.e., a reading log with comments, literature response journal); however,

the majority of these were coded only as reading because writing was not the focus of instruction, simply the mode of response.

Evaluation criteria. Each site had developed and used its own construct-centered evaluation system to accompany its outcomes and portfolio model (Messick, 1994). The *Bellevue evaluation rubric* was designed by participating teachers around its major Student Learning Outcomes—Reading Ownership, Reading Ability, Reading Variety, Reading Strategies, Writing Ownership, Writing Ability, Writing Process, and Self-Reflection/Evaluation. It included descriptors of performance for each outcome for ratings of 1, 3, and 5, although raters could assign scores in between (Valencia, 1996). This could be considered a "high inference" system. Descriptions do not specify the form of types of information that must be found in the portfolio. Instead, there are general criteria associated with each particular outcome and performance level. An example for reading comprehension (Reading Ability) is presented in Table 3. The evaluator looked for the required common tools in the Bellevue portfolios such as reading logs and questions, reading summaries, and reading response journals, although these tools might have provided evidence on other outcomes as well. Additional evidence for reading comprehension might have been found in the form of self-reports, running records, reading activities, assignments, or anecdotal records. The evaluator wrote down supporting evidence and conclusions, compared the evidence with the rubric, and then assigned a rating of 1–5 or M for "missing." This procedure was followed for each outcome. In the Bellevue system, the specific descriptors associated with a rating of 3 were designed to represent typical performance for each outcome for students at the designated grade level. By using both descriptors and grade-level designation, teachers felt they would be able to provide useful information to many different audiences.

The *KEEP evaluation system* was designed by KEEP consultants, and similar to the Bellevue system, it was designed around its four major literacy outcomes: Reading Ownership, Reading Process, Writing Ownership, and Writing Process. Within each of these outcomes is a list of specific benchmarks, indicating the kind of evidence that should be found in the portfolios (Asam et al., 1993). There are between 7 and 21 benchmarks for each literacy outcome at the primary level, and 9 to 28 benchmarks for each at the intermediate level. The number of benchmarks increases by grade level. Specific evidence is required for each benchmark. For example, writing samples, including drafts and published versions, are required for benchmarks under Writing Process. This could be considered a "low inference" system compared with the Bellevue system, since it is quite specific about what counts as evidence. An example for comprehension is presented in Table 3. When examining the portfolio, the evaluator had to find evidence for each benchmark. For example, for "writes personal responses to literature," there needed to be copies of pages from a literature log. Some benchmarks were covered by the same piece of evidence. For example, a written story summary might have served as evidence for the student's ability to comprehend and write about the theme, as well as

TABLE 3

Comparison Across Three Evaluation Criteria for Reading Comprehension (Intermediate Grade)

Bellevue	KEEP	Common
(5) Personal response; synthesis; coherence; theme, major concepts; significant details; coherent/concise summary; applies to prior knowledge; grade-level text or above	*Written Response: Aesthetic* •Writes personal responses to literature •Comprehends and writes about theme/author's message •Applies/connects theme to own life/experiences •Makes connections among different works of literature	(5) Coherence; personal response; theme/big idea; story elements; significant details; connections; reads different types of material; has a wide range of reading interests (more than 3 types); grade-level text or above
(3) Some personal response; attempt to synthesize; main idea or problem; some details; literal focus; logical sequence; grade-level text or above	•Applies/connects content text information to own life/ experiences *Written Response: Efferent* •Reads nonfiction and shows understanding of content	(4) Not as much coherence or insight, but still has a personal response; theme/big idea; story elements; significant details; connections; reads more than 3 types of material; grade-level text or above
(1) Limited response; summaries are retellings; sketchy; basic facts; misinformation or misunderstandings; grade-level text or below	•Writes summary that includes story elements •Uses clear, meaningful language to express ideas in written responses or summaries	(3) Retelling present or poorly written summary; surface information; sketchy; mostly literal understanding; story elements; reads 2–3 types of material; grade-level text or above
(M) No evidence in any of the artifacts; evidence missing	•Reads different genres of fiction and shows understanding of genre characteristics •Understands elements of author's craft	
	Research Strategies •Obtains facts and ideas from a variety of informational texts •Uses a variety of reference materials •Uses graphic organizers •Uses a variety of library resources •Takes notes •Writes research report synthesizing information from multiple sources •Publishes research report of equivalent product •Uses a variety of outside resources	(2) Incomplete retelling; little coherence; details; reads 1–2 types of material; grade-level text or below (1) Minimal response; little or no reading; grade-level text or below (M) No evidence in any of the artifacts; evidence missing

knowledge of story elements. However, the evaluator of a third grader's portfolio had to find evidence for all 19 benchmarks. On the basis of the evidence, the evaluator rated the student S (satisfactory, grade-level performance), D (developing, below grade-level performance), or M (missing, no evidence available) for each specific benchmark. For

benchmarks that depended on the story summary as evidence, the evaluator referred to anchor pieces serving as examples of performance at or below grade level. If, after examining all the evidence, the student received an S on all benchmarks, the evaluator assigned an overall rating of "at grade level." If the student received one or more *D*s, the rating given was "below grade level."

The *Common evaluation system* was developed collaboratively by the teachers at the second cross-site meeting. It more closely resembled Bellevue's system than KEEP's; it was oriented to just seven outcomes, rather than to a whole series of benchmarks. The outcomes were Reading Product, Reading Process, Reading Enjoyment, Writing Product, Writing Process, Writing Enjoyment, and Self-Evaluation/Reflection. Like the Bellevue system, the Common system used a descriptive 6-point scale, with a rating of 3 describing typical grade-level performance. The descriptors for the points on the scale were somewhat more specific than in the Bellevue system, but not as specific as in the KEEP system. Thus, the Common evaluation system required more inference than the KEEP system but less than the Bellevue system.

The treatment of reading comprehension in the Common evaluation system is similar to that in the Bellevue system, focusing on a broader outcome—Reading Product. However, the teachers also liked the specificity provided by the benchmarks in the KEEP system. They felt that rating portfolio evidence would be easier if descriptors could be made more precise and detailed than in the Bellevue system. For this reason, they began developing descriptors for ratings of 1 through 5 for the Common evaluation system (see Table 3). The teachers' intention was to integrate some of the KEEP benchmarks into the descriptors, but time did not permit them to accomplish this task. However, they spent considerable time discussing how concepts listed under the benchmarks would fit within each of the outcomes. So, although the written documentation was not fully developed (see Table 3), their discussion helped them clarify and specify their understanding of the criteria. They also suggested that it would be useful to specify required common tools to be used in the future for the new evaluation system.

Interrater agreement was calculated among evaluators across sites. Using the KEEP nominal scale of S, D, and M, agreement among raters required an exact match. Using the Bellevue and Common scales, ratings that were within 1 point were considered matches as were matches in "missing" designations.

Professional development. Teachers' professional development was examined qualitatively by examining the information provided in the individual interviews conducted at the beginning and end of the study. In addition, audiotapes made at the meetings provided evidence of what teachers were learning. Audiotapes were made of all whole-group discussions, portfolio orientation sessions, and debriefing discussions after teachers' observed each other's classrooms. Records of the contents of portfolios teachers submitted at the first and second meetings served as another source of information about changes in the teachers' understandings.

Table 4
Content Analysis of Portfolios

	Primary		Intermediate	
	Bellevue	**KEEP**	**Bellevue**	**KEEP**
Average total number of pieces[1]	18	18	22	23
Disciplinary Focus				
reading	13	7	6	2
writing	8	8	8	11
other subjects	4	2	5	1
Type of Artifact				
student work	16	15	20	14
anecdotal notes/checklists	1	4	0	6
other (photo, drawing, audiotape)	2	1	1	2
student-selected/self-assessment/ reflection	8	5	10	6
parent entry/comment	1	0	0	1

[1] All numbers are averages across the portfolios evaluated. Each portfolio artifact was coded according to the variables listed below. As discussed in "Data Sources," some artifacts were double coded. Therefore, column totals do not equal average total number of pieces.

Findings and Discussion

Portfolio Content Analyses

The contents of the portfolios were used to determine if they included authentic, high quality samples of the curriculum outcomes, and to compare the types of artifacts across sites and over time. There had been no a priori agreement about the portfolio evidence nor had there been any contact between teachers across sites before the first meeting. Table 4 presents a summary of the average number of types of artifacts included in all 8 portfolios analyzed at

the first meeting. As noted above, artifacts were coded according to several criteria, which resulted in double coding of some artifacts.

The analysis revealed surprising similarity between Bellevue and KEEP portfolios in terms of the average number of pieces collected in each portfolio between September and January. However the distribution of pieces across sites varied considerably and was consistent with the emphasis at each site. For example, Bellevue portfolios included many more reading pieces than KEEP portfolios, reflecting attention to collecting both reading and writing portfolio artifacts in the Bellevue portfolios. On the other hand, there were only a few more writing pieces in KEEP portfolios than Bellevue's even though 3 of the 4 KEEP teachers emphasized collecting evidence from their writers' workshop. Teachers confirmed that it was easier to collect writing evidence than reading evidence, and suggested that without deliberate attention to collecting reading artifacts, it was difficult to represent reading in the portfolios. Consequently, an emphasis on reading in the portfolios resulted in more reading evidence, but a corresponding emphasis on writing in portfolios did not produce more writing evidence.

Several different types of artifacts were found in the portfolios. The most common, however, was student work. Within this category, student-generated written work predominated. For example, there were many samples of students' original writing such as stories, reports, responses to what they were reading, and daily journals. Often there were rough drafts and final published pieces. There were also several surveys and reading logs. Only a few, 2 out of 81, of the artifacts were lower level, fill-in-the-blank types of activities. The overwhelming majority of student work in these portfolios represented authentic instances of students' literacy learning; artifacts of student performance resembled meaningful reading and writing behaviors and tasks. For example, there were pen pal letters, selected pages from learning journals, book logs, original poetry, and reading and writing interest surveys from the beginning of the school year. One teacher remarked,

> I think one of the reasons that our project worked is because we were really looking at . . . very authentic evidence of kids' work. There weren't any worksheets in there, and I'm not just panning worksheets, . . . but I felt like we were looking at authentic evidence of what kids could produce.

There were slightly more student-selected pieces, goal setting artifacts, and portfolio visits in Bellevue portfolios, consistent with their portfolio model and Self-Reflection outcome. On the other hand, KEEP portfolios included substantially more teacher observation and anecdotal data. This is most likely because several KEEP benchmarks required completion of observation checklists and one required running records.

There was more consistency in both disciplinary focus and types of artifacts among portfolios from the same classroom than among portfolios from the same sites; we could easily identify portfolios from the same classroom. For example, one class had spent several weeks researching and writing state reports. As a result, the reading and writing portfolio artifacts in those portfolios were distinctive from other classes. Some of these artifacts

served as required pieces, others were selected by the teacher or student as personal entries. Although teachers included items according to their specific site guidelines, they also included a variety of artifacts they, or their students, wanted to include. Even with this individuality, there was still a marked similarity among portfolios from the same site; Bellevue portfolios looked more like other Bellevue portfolios than like KEEP portfolios. Bellevue's portfolios were distinctive for their "common tools" (i.e., reading summaries, book logs, entry slips, portfolio visits), and KEEP portfolios were distinctive for their reading and writing checklists.

While reviewing each other's portfolios at the first cross-site meeting, teachers spontaneously discussed types of evidence that were new or interesting to them. For example, Bellevue teachers noted how KEEP teachers documented the planning stage of writing and how they used the benchmarks to help students self-evaluate their writing. Similarly, KEEP teachers were interested in Bellevue's common tools, use of entry slips, and the ways teachers documented reading as well as writing. A review of portfolio contents at the second cross-site meeting revealed that teachers from both sites had incorporated some of the teaching strategies and portfolio artifacts they had learned from each other. Bellevue portfolios contained more evidence of planning for writing, using rubrics with students, and making intertextual connections. KEEP portfolios contained more entry slips, reflections about learning, and information related to students' reading. In some cases, these new artifacts reflected an effort to document student learning that teachers reported they had observed but had not included in a portfolio. For example, most teachers reported holding reading conferences with students and having them plan for writing, but they had not focused on including conference notes or planning notes in the portfolios. In other cases, the new artifacts reflected a new instructional emphasis teachers had learned during observation, discussion, and evaluation. For example, at the first meeting, several teachers discussed the importance of having students make intertextual connections during literature discussions. Other teachers found these ideas new and intriguing. By the second meeting, several teachers had tried to help students make more meaningful intertextual connections and had included that evidence in their portfolios.

We should note, however, that teachers were as struck by the similarities in their philosophies and instructional emphasis as by the differences in the portfolio contents. As one Bellevue teacher put it, the contents of the KEEP portfolios

> are somewhat different from ours, but they ask a lot of the same things . . . What the student produces is very similar. Our portfolios might look a little different, but you get the essence of the child, I think, in both of them.

A KEEP teacher noted:

> It was so neat to find that the Bellevue teachers and KEEP teachers shared many common goals in literacy. We found many similar successes as well as challenges. How wonderful that we could examine portfolios separately and still come up with common agreement and understanding of an individual through [the] portfolio.

Table 5
Interrater Agreement: Site by Evaluation Criteria

	KEEP Portfolios	Bellevue Portfolios	Total
KEEP Criteria	80	74	77
Bellevue Criteria	81	89	85
Common Criteria	87	97	92

With a shared understanding of literacy learning and teaching behind them, and a familiarity with one another's portfolio artifacts, these teachers could see the commonality in learning outcomes and uniqueness of individual children in portfolios from different sites and classrooms.

Overall, the content analyses suggest that portfolios contained high quality, authentic samples and records of students' reading and writing. They reflected the underlying outcomes and portfolio models of each site as well as the emphasis of individual teachers. However, evidence of some outcomes and some types of work, particularly reading outcomes and discussions, were apparently difficult to document using a portfolio, resulting in limited information about particular outcomes. Nevertheless, the process of reviewing and discussing portfolio contents provided teachers with ideas about both assessment and instruction that carried over into their own classrooms and their students' portfolios.

Evaluation Process and Results

Qualitative and quantitative data from the evaluation sessions provide insights about the feasibility and desirability of cross-site evaluation of classroom-based portfolios. Results indicate that teachers from different sites were able to learn and apply their own and other evaluation systems. Interrater agreement, averaged across outcomes, increased from the KEEP, to the Bellevue, to the Common criteria for portfolios from both sites (see Table 5). The relatively strong interrater agreement suggests promising possibilities for reporting results for groups of students across sites. Such agreement is remarkable because the portfolios did not contain similar prespecified pieces; they were developed to be most useful at the local district and individual classroom level. As a result, there was substantial variability in the contents across portfolios from different sites. Additionally, because most of the pieces in the portfolios were complex authentic examples of reading and writing, individual pieces often contained evidence of several outcomes or benchmarks. For example, a reading response journal might contain evidence of reading ownership, writing ownership, reading comprehension, and writing process. The evaluator would need to review each piece carefully so that all possible evidence was identified within a particular artifact. Furthermore, individual

portfolio artifacts did not receive ratings; instead, the entire contents of each portfolio was reviewed and then a rating given for each reading and each writing outcome or benchmark. Evaluators reviewed 12 to 25 pieces within each portfolio. Finally, the teachers had only a brief training for evaluation, and then spent only 30 min rating each portfolio for all the outcomes or benchmarks.

Findings from the evaluation process also provide a perspective on one specific aspect of validity—content or construct underrepresentation (Linn et al., 1991; Messick, 1994). Although most proponents suggest that portfolios provide valid assessment of literacy learning, three assumptions underlie this belief. First, it's assumed that all identified outcomes or benchmarks are adequately documented in the portfolio. Second, the evaluation criteria used to evaluate the portfolios must be aligned with these outcomes or benchmarks. Third, it's anticipated that teachers can agree on the evidence in a portfolio that can be used to evaluate each specific outcome or benchmark. These are difficult conditions to achieve when portfolios are generated and evaluated within a single site—significantly more difficult when portfolios are generated and evaluated across different sites. In cross-site evaluation, it is important to determine if there is a fit between portfolio evidence and the evaluation criteria used within and across sites. Additionally, teachers must be able to interpret different types of complex portfolio evidence produced in different contexts to fit site-specific outcomes or benchmarks. To address these issues of fit, we examined the extent to which the actual portfolio artifacts were judged to represent evidence of the outcomes and benchmarks included in the various evaluation criteria. Specifically, we looked for missing data to determine if the portfolios from each site provided sufficient evidence to make judgments about student performance. If teachers cannot find or interpret artifacts needed to apply specific evaluation criteria, then the validity of the evaluation should be questioned.

Figure 1 presents the percent of missing data recorded for portfolios from each site according to each evaluation criteria. As teachers rated, they looked for evidence of specific outcomes or benchmarks in each portfolio. If an individual rater determined that there was no evidence in any of the artifacts, she recorded "missing." The percentage depicted in Figure 1 represents the percent of all the ratings assigned to each portfolio that were designated "missing." There was a substantial amount of missing evidence recorded by the teachers when using the KEEP evaluation criteria, less for the Bellevue criteria, and slightly less for the Common criteria.

In-depth analysis suggests that the missing data for the KEEP evaluation system was a result of several factors. First, the KEEP criteria required very specific indicators of performance and specific types of information (e.g., running records, observation checklists); if specific types of evidence were not found, "missing" was recorded. This could be considered a problem of— grain size," that is, needing to find very particular indicators of performance. Second, three KEEP teachers did not systematically collect portfolio information on reading and one did not collect information on writing. Although the KEEP teachers taught

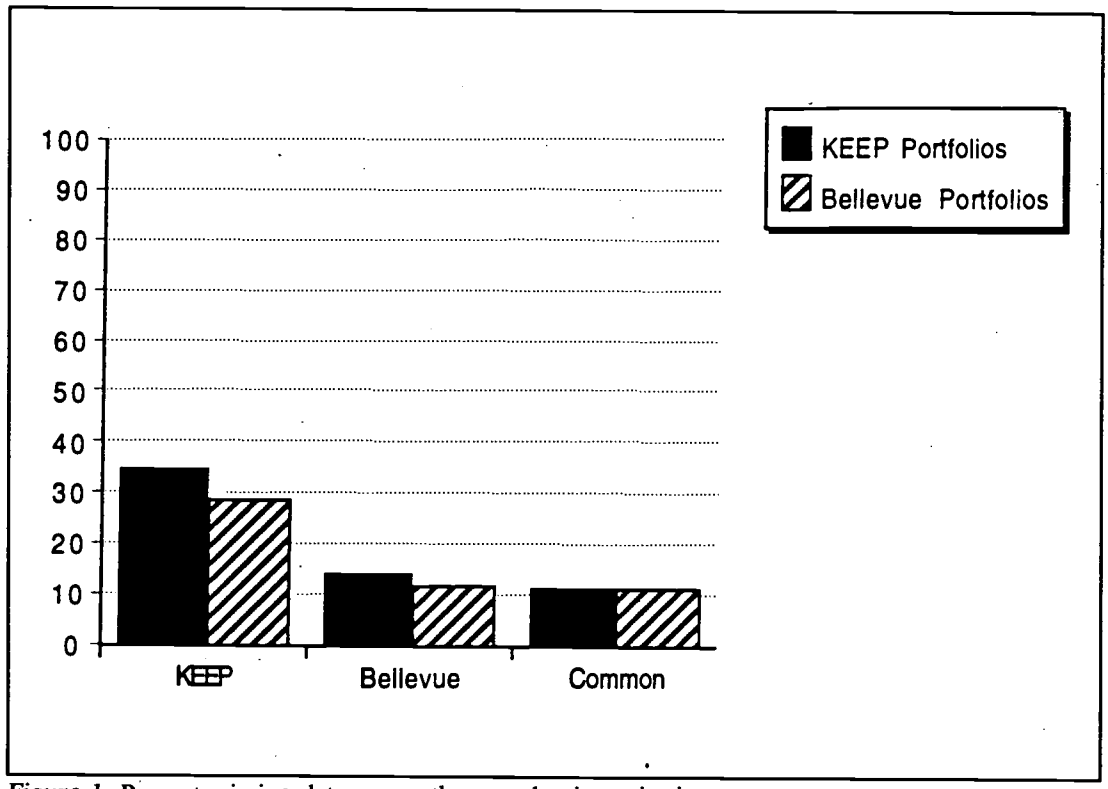


Figure 1. Percent missing data across three evaluation criteria.

both reading and writing, they did not try to focus portfolio information on both. Consequently, there was a significant amount of missing data for reading in the KEEP portfolios when evaluating with the KEEP criteria. On the other hand, there was also a substantial amount of missing data in the Bellevue portfolios when the KEEP criteria were applied. This was because Bellevue teachers did not include the very specific kinds of evidence required for some of the KEEP benchmarks (e.g., running records, preplanning for writing, writing in multiple genres).

In contrast, there was relatively little missing data found when the Bellevue evaluation criteria were applied; this was most likely a reflection of the more global nature of the Bellevue descriptors. Missing evidence in KEEP portfolios was predominantly in the area of student self-reflection, which was not a KEEP benchmark. KEEP teachers had not emphasized student self-reflection in their teaching or in their assessment. As a result, there were few, if any, artifacts of self-reflection in the KEEP portfolios. There were also some missing data in the Bellevue portfolios in the area of Reading Strategies. Bellevue teachers had difficulty documenting this outcome even though it was part of their curriculum.

The Common evaluation system, which was a synthesis of the KEEP and Bellevue criteria

(adopting the global outcomes of the Bellevue rubric with some of the specificity of KEEP benchmarks), produced missing data resembling that identified for the Bellevue rubric. Missing data were found primarily in areas of self-reflection, writing process, and reading process. These areas replicated the missing data identified when the KEEP and Bellevue criteria were applied.

These findings suggest that missing data are not only a problem when applying an "outside" evaluation criteria to local portfolios but when applying the "inside," locally developed criteria as well. If teachers or students are unfamiliar with the evaluation criteria ahead of time, they may not select work required by the criteria. This does not necessarily mean that outcomes or benchmarks were not taught or learned, or that the portfolios could not or would not include this evidence. It simply suggests that there was not a perceived need to include it. For example, after reviewing and evaluating KEEP portfolios, one Bellevue teacher commented that she wanted to try to document small group discussions and research projects in her portfolios. Although these were already a part of her instructional program, there was no evidence in her students' portfolios to demonstrate these abilities. Alternatively, it is possible that evidence was missing because a particular outcome was NOT part of the class curriculum. An example would be self-reflection for the KEEP portfolios.

The nature of these missing data suggests that a specific evaluation system used with a classroom-based portfolio may not provide a valid measure of a particular child or of a particular classroom (e.g., no information placed in the portfolio; using evaluation criteria that were not intended for a particular site). To remedy this might require a fairly minor adaptation, for example, requesting that teachers include specific artifacts typically found in their classrooms. Both the KEEP and Bellevue teachers assured us this was possible and desirable. Alternatively, it might require a more dramatic intervention—a change in the curriculum (outcomes/benchmarks), instructional emphasis, or revision of the evaluation criteria to align it with the actual curriculum.

Although there was a considerable amount of missing evidence, teachers across sites generally were able to agree on the presence or absence of evidence. On average, teachers reached 94% interrater agreement about whether evidence for each outcome could be found in the artifacts or whether evidence was missing. They were able to examine many different types of complex portfolio artifacts generated in different classrooms, and agree on what counted as evidence of particular learning outcomes or benchmarks. It appears that the teachers shared a strong knowledge base in literacy that enabled them to interpret student work consistently. There were few instances (less than 6%) of teachers from one site finding sufficient evidence to assign a rating and other teachers finding no evidence.

Cross-site evaluation of portfolios also provides insight about teachers' abilities to use a different set of outcomes and criteria to evaluate their own and others' portfolios. In addition to the complexities outlined above, teachers had to step outside their own portfolio models, artifacts, instructional strategies, evaluation systems, and personal knowledge of

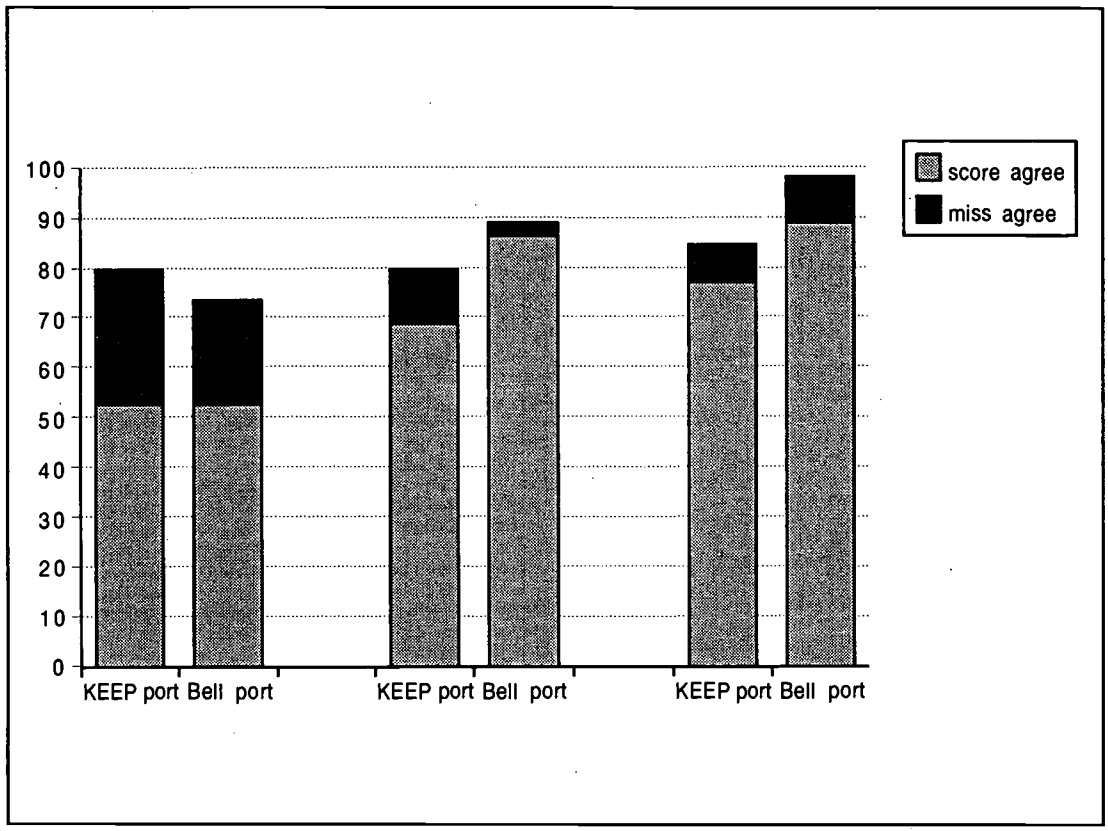


Figure 2. Interrater agreement across three evaluation criteria.

their students to evaluate the portfolio artifacts. We addressed this aspect of evaluation by examining if teachers' ratings of portfolios from their own site were different from ratings given by the "outside" teachers.

Our analysis suggests that teachers did not have a bias when interpreting and rating portfolios. Bellevue teachers and KEEP teachers did not consistently rate their own portfolios higher or lower than the teachers from the other site. KEEP teachers never rated their portfolios higher than Bellevue teachers and Bellevue teachers rated their own portfolios higher than KEEP teachers less than 1% of the time. All the teachers appeared to be able to use the portfolio information, adjust their expectations to the criteria being used, and apply it equally to their own and others' portfolios. In light of concern about fairness of new assessments (Linn et al., 1991), this finding is especially promising. Not only were the students from very different cultural backgrounds, but the teachers (raters) were as well.

Overall, the evaluation data suggest that these teachers learned to rate portfolios with a high degree of interrater agreement using several different, but philosophically similar, evaluation systems. Several factors most likely

contributed to this consistency. First, teachers shared common conceptual understandings about literacy learning that had been developed for 4 years at each site before this project began. These understandings were further enhanced through the classroom visits, cross-site portfolio discussions, and the evaluation process that were part of this study. This common knowledge was reflected in the teachers' abilities to interpret portfolio work from their own site and the other site. One teacher commented:

> I think what surprised me most was how easily we could communicate with the teachers about portfolios. . . . We almost immediately got into good discussions about how the evidence was collected. I learned that we could talk to teachers from a very different setting about portfolios and come up with much agreement about the students' skills and performances. The KEEP teachers saw things in my portfolios that transcended the tools and "spoke" of the student.

Integral to this professional collegiality was the low stakes nature of the project. Teachers felt fortunate to be part of this working group, were generous with their colleagues, and eager to learn. One Bellevue teacher noted that,

> With the KEEP (teachers), I think there was more opportunity to really look at what people were doing and then say, "Gee, I think I need to be doing that" whereas in a lot of our own meetings, . . . it's sometimes been competitive.

A second explanation for strong interrater agreement might be the inclusion of missing data. Figure 2 depicts the difference in interrater agreement when missing agreements are added or deleted from the calculations. Obviously, interrater agreement would have been substantially lower, especially using the KEEP rubric, had we not counted "missing-missing" agreements into our calculations. However, most evaluation systems do not use a "missing" category, but rather rate no evidence or minimal evidence with the lowest value on the evaluation scale. We believe that the ability to interpret complex work in portfolios, and to see (or not see) evidence of outcomes are important issues for portfolio evaluation and professional development. This is especially true if portfolios are locally defined but evaluated across sites; the complexity and variability will always be present to complicate interpretation.

We also believe missing data shed light on the alignment of curriculum and assessment. A substantial amount of missing evidence suggests a lack of alignment, an underrepresentation of the targeted outcomes (Messick, 1994). Consequently, the validity of the assessment would be questioned. Our experience suggests, however, that teachers can become more focused on collecting portfolio evidence when they know what is expected and believe that evidence would be useful to them as well as others. Nevertheless, even with the best intentions, we found that some types of evidence were difficult for teachers to document.

Finally, the increase in interrater agreement across evaluation criteria was probably somewhat influenced by the progression from a very specific evaluation system (KEEP) to more global criteria (Bellevue), to the collaborative development of the Common criteria which teachers felt combined the best of both systems. In addition, although 4 months separated

the two meetings, and teachers did not evaluate portfolios at their local sites during this time, they were interested and motivated to prepare for the next cross-site meeting. This, in turn, focused their attention on the portfolio contents, the outcomes and benchmarks, and the evaluation process, which might have improved their interrater agreement over time.

The evaluation process was judged by teachers to be a valuable part of portfolio implementation. Their comments suggest that collaboratively learning to identify evidence and then rate that evidence, provided insight about assessment and instruction that they might otherwise not have gained. Because there were low stakes placed on the scores and high priority placed on professional development, teachers were not intimidated by the evaluation process or defensive about the results. They were able to take full advantage of the opportunities. They found the experience to be engaging, interesting, and professionally useful. Said one teacher,

> I can't remember when an experience like this gave me as much to think about in terms of what I could adopt and adapt. The chance to re-examine what we had developed by looking in a very similar but different mirror is a very unusual opportunity and (one) to be highly prized.

Professional Development

A more detailed examination of professional development issues is made possible by looking closely at two teachers, one from Bellevue and one from KEEP. These two were selected because they typify the eight teachers in the study. They were articulate and eager to talk about their experiences. Through their words and experiences, we gain insights that are evident in the protocols of the other participating teachers. We analyzed the statements made by these teachers during interviews and meetings and looked at their students' portfolios for evidence of the influence of the project upon their professional development. The teachers' names are used with their permission.

Sue Bradley, who taught a third- through fifth-grade multiage classroom, had 12 years of teaching experience at the time of the project and had a strong background in the use of portfolios. Sue was an original member of the Bellevue portfolio group, had used portfolios in her classroom for 4 years, and had been one of the district's representatives to the New Standards Project. When Sue first observed the classroom of Nora Okamoto, a KEEP fifth-grade teacher, she immediately noticed similarities to her own teaching. "For example, writers' workshop, literature circles, response journals were all things that I instantly recognized in Nora's room." These similarities allowed her to focus on specific details of Nora's instruction, such as "how she paced her day, what she asked of students in and out of class, and visible signs around the room of the accountability that she had established for her students." The similarities in philosophy and instruction made it easy for Sue to communicate with Nora and the other KEEP teachers about portfolios. "There was little lead time needed to build a common understanding, and instead we almost immediately got into good discussions about how the evidence was collected."

After her visit to KEEP, Sue noted that she wanted to try more of the modeling of reading

and writing with her students that she had seen Nora do. In the classroom of another KEEP teacher, she was impressed by the student-led literature discussions and the emphasis on quality responses. She remarked on the KEEP teachers' constant use of the terms of the KEEP reading and writing benchmarks (e.g., author's craft, reading/writing connections, reading and writing in different genres), which gave students the language for communicating with teachers about their progress. Finally, she thought it was a good idea to involve students in assisting with record keeping.

Involvement in portfolio evaluation and development of the Common evaluation rubric, with the complex discussions that these entailed, broadened Sue's understanding of portfolios. She stated:

> I think it's a healthy process because any time you have to teach somebody what you mean, it is like peer teaching in a classroom. It's not just giving, it's a very reciprocal situation. So when you have to explain what you meant in a certain situation to someone else, you're then clarifying your own thinking.

The discussions helped Sue to see the significance of various artifacts within another teacher's portfolio. She stated, "Because I had conversations with her (Nora), sat and wrestled with scores with her, I gained more of an understanding about what that chart (a portfolio artifact) represents."

At the conclusion of the project, Sue noted that she had started being more specific with her students about the kinds of evidence needed to document their involvement with the writing process. She described herself as "much more religious about looking for evidence of pre-planning writing" to the point where she required her students to write out their plans instead of making this step optional. She had her students evaluate their own writing using an evaluation rubric, as a means of teaching them about standards for literacy performance. She also had become more aware in her own classroom of the quality of students' responses during literature discussions.

The portfolio content analysis served to verify that Sue had succeeded in making several changes. For example, one student's portfolio contained a plan for writing and a penpal letter (to a student in Nora's class) with a prewriting web. Clearly, this student had become aware of gathering evidence for planning her writing. This portfolio included a piece of explanatory writing with two student-scored rubrics and a checklist on which the student had evaluated her progress as a writer. These artifacts indicated that the student was involved in activities to promote awareness of standards for literacy performance. The portfolio also contained the notes of a reading conference, showing Sue's concern with the quality of the student's responses to literature.

As noted above, Sue's counterpart at KEEP was Nora Okamoto. Nora had been teaching for 21 years and associated with KEEP for 4 years. At the time of the project, she had used portfolios in her classrooms for 2 years, but only as part of her writers' workshop, not her readers' workshop. In the first year, following the KEEP approach, she had familiarized the students with the fifth-grade benchmarks for writing. Then she had gone through the students' portfolios and tagged pieces showing evidence that the students had met the various

benchmarks, a documentation procedure required in KEEP's monitoring of student achievement. During the second year, Nora decided to turn the responsibility for selecting and tagging evidence over to the students. She felt this approach had worked to address the only weakness she saw in using portfolio assessment: that it could be extremely time-consuming. Before meeting the Bellevue teachers, Nora indicated that she was in the process of developing a system to assess her students' progress in reading to parallel the system she had worked out for writing.

Contact with the Bellevue teachers accelerated Nora's thinking about how to expand her portfolios to encompass reading as well as writing. Nora found that immersing herself in another portfolio system "afforded me the opportunity to see a wide range of data collection tools that work well in assessing student achievement and growth." By studying the Bellevue portfolios, she gained many specific ideas about how to document growth in reading. Before her visit to Bellevue, Nora had not seen a multiage classroom or a school building organized into pods to encourage team teaching. However, like Sue, she was able to see beyond the differences in classroom and cultural contexts, noting Sue's students' involvement in many of the same activities she used in her own classroom, including literature circles, independent reading, and reading journals.

Nora indicated that involvement in portfolio evaluation and development of the Common evaluation rubric had "broadened [her] perspective on portfolio contents and organization." She gained confidence in her ability to understand descriptors and evaluate individual students' portfolios. She was impressed by the group's ability to reach agreement about the criteria for judging students' competence in reading and writing.

At the conclusion of the project, Nora observed that she had become aware of the importance of students' self-evaluation and reflection and regular portfolio visits, central features of the Bellevue portfolio system. She noted:

> I've expanded on the student entry tags on the writing process steps to include comments about specific skills, and I now have students evaluate writing projects upon completion (self-evaluation and reflection). I plan to use colored entry slips and the portfolio visit questionnaire as part of the portfolio next year. The entry slips were helpful in understanding the significance of the work and why it was selected for the portfolio. Quarterly portfolio visits provided insight into students' thinking about their overall progress in that quarter.

Nora concluded that "participation in this project has had a positive impact on my learning and growth as a classroom teacher." She added that she had gained "methods to try, tools to use, and a better understanding of portfolios as a way to integrate teaching, learning, and assessment."

A content analysis of a sample of Nora's portfolios at the end of the school year indicated that she had already succeeded in making several changes. At the end of the second quarter, after her first meeting with the Bellevue teachers, Nora had her students add the following to their portfolios: a page from their reading journals; a web showing efferent and aesthetic responses to literature; and activity

sheets showing vocabulary development, story sequencing, and an analysis of story parts. These items were evident in students' portfolios in the fourth quarter. Entry slips similar to those used in Bellevue were placed on the items, indicating what students thought the items showed about their progress as readers and writers.

As these case examples indicate, the teachers in this project gained specific ideas for improving portfolio assessment in their classrooms. Their study, rating, and discussions of each other's portfolios heightened their concerns about documenting students' literacy learning in as complete and sensitive a manner as possible. Teachers seemed to be adding to their notions of what should be included in a portfolio and broadening their thinking about important dimensions of student performance. Their approach seemed to be additive; they did not discard any of the dimensions recognized as important in their own portfolio system, but added on dimensions highlighted in the other site's portfolios (for example, self-evaluation in the case of Bellevue and the planning of writing in the case of KEEP). In other words, they appeared to be expanding the criteria by which they assessed students' literacy learning, knowing that their instruction would have to be adjusted to meet these additional criteria. As Frederikson and Collins (1989) suggest, developing an evaluation system and learning to apply it can make the critical traits in performance clear to teachers as well as students.

Understanding a Portfolio Assessment System: The Components and Conditions

Any discussion of this case study must take into account the naturally occurring context, both for its best-case perspective and for the realities of classroom-based research. Portfolios had been well-established, integral elements of these teachers' classrooms and, as such, they reflected the experience, knowledge, commitment, professional development, idiosyncrasies, and individualization each teacher brought to her own classroom and to the cross-site collaboration. Such is the nature of all classroom-based assessment and, indeed, all instruction. We found that rather than serving simply as assessment tools, effective portfolios were actually part of a complex and interactive portfolio system that served to support both teaching and learning. Our purpose, therefore, is not to generalize to other projects, list the specifications for portfolio contents, or suggest evaluation criteria, but to highlight the essential components and the internal and external conditions which create a successful portfolio assessment system.

The Components

The three components described in this study—portfolio evidence, the evaluation process, and professional development—are at the heart of such a portfolio system. All must all be in place if portfolios are to improve instruction, student achievement and classroom-based assessment. Each component has a profound influence on the others and on the overall effectiveness of the system (see Figure 3). A brief description of the interaction among these components follows.

Interaction of portfolio evidence and professional development. Our data and experiences suggest that as teachers closely examine stu-

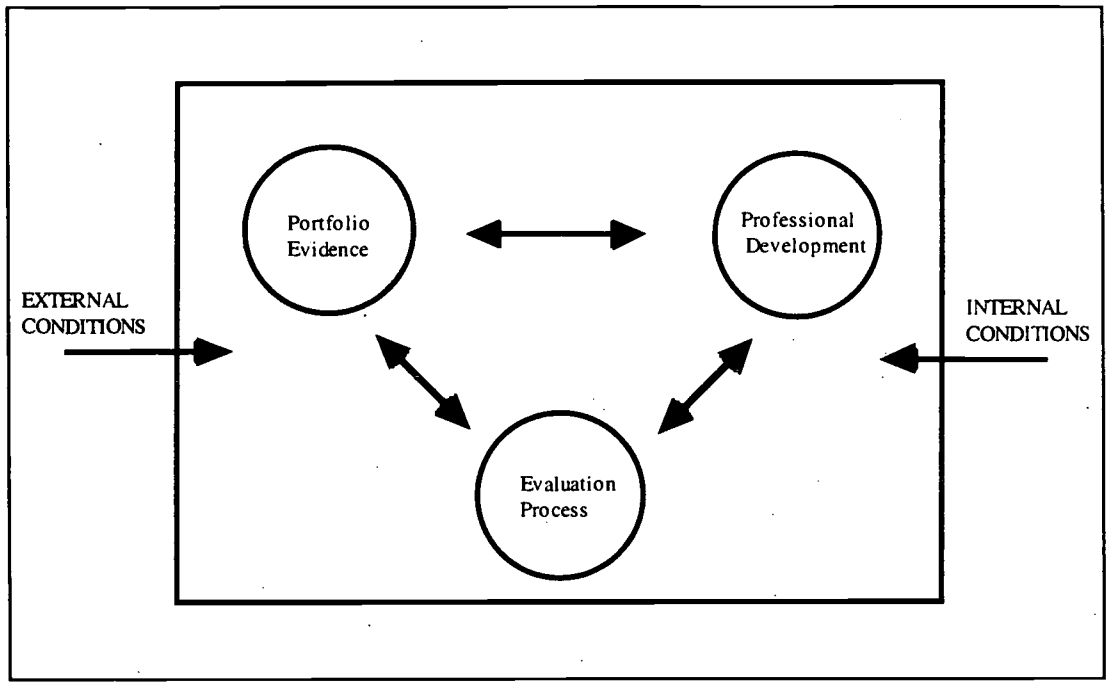


Figure 3. Model of effective portfolio system.

dents' work from their own and other classrooms, they clarify important learning outcomes and learn to interpret student performance based on multiple forms of evidence. Working collaboratively, discussing portfolio artifacts, encourages teachers to reexamine their knowledge, assumptions, and misconceptions about teaching and assessment practices. Conversely, teachers' understandings of curriculum and instruction are reflected in portfolio evidence. We can see their strengths and weaknesses, their priorities, and the opportunities they provide for their students. As teachers grow and change through professional development, their portfolio evidence begins to change as well. This slow, iterative process is likely to produce meaningful, sustainable changes in both portfolio evidence and in underlying classroom instruction.

Interaction of portfolio evidence with the evaluation process. Using portfolios for evaluation forces teachers to rely on evidence for their decisions. It grounds evaluation in concrete data. Teachers also learn that students cannot be evaluated for learning that has not been documented or that has not been taught. When portfolio evidence is inadequate or unavailable, the evaluation process may be complicated and the results may be misleading or subject to question. Results will differ dramatically depending on the interpretation, sources, and amount of portfolio evidence that is missing. At the same time, the evaluation process can reveal gaps in portfolio evidence that, in turn, reflect gaps in teaching and class-

room assessment. As a result, portfolio contents and related instruction are likely to change to match the evaluation criteria. It is clear, however, that genuine changes in portfolio evidence occur only when teachers find portfolios useful in their own classrooms.

Interaction between the evaluation process and professional development. The evaluation component is absent from many local portfolio projects. Many reject it because of negative associations with standardized test scores and psychometric difficulties of portfolio evaluation. However, when teachers are involved in the development of evaluation criteria and the actual evaluation process, the nature of the interaction between evaluation and professional development changes from an antagonistic one to a symbiotic one. The evaluation process becomes a positive and valued vehicle for teachers to examine their instruction and expectations for children. Evaluation of portfolios anchors teachers' expectations beyond their local classrooms, and it reinforces the importance of making criteria clear to students. In addition, with practice, teachers become better evaluators, more able to identify evidence in complex student work and to make reliable judgments about the quality of that work. And, as noted above, once teachers are familiar with the evaluation criteria and types of evidence that are useful, they can adjust their portfolios and instruction accordingly.

Teachers who have developed a strong knowledge base in teaching and learning have little difficulty applying evaluation criteria to portfolios from their own and others sites; those who are less well-grounded would likely experience difficulty. Furthermore, teachers' knowledge and professional development experiences influence their participation in the evaluation process. Experienced portfolio teachers use the process to ask questions, learn new strategies, and support their colleagues. They enrich the evaluation process and take advantage of the opportunity for professional growth.

Internal and External Conditions

Evidently, a push or pull on one component within this system affects the others. In addition, the nature of these interactions is influenced by the internal and external contexts in which they exist. Bird (1988) reminds us that "the potential of portfolio procedures depends as much on the political, organizational, and professional settings in which they are used as on anything about the procedures themselves" (p. 2).

Supportive internal conditions are essential to sustaining an effective portfolio system. In this study, the local conditions within each site supported the emerging portfolio systems. Stakes were low for teachers; district interest, support and commitment high. Local, long-term professional development encouraged gradual implementation of portfolios and development of an evaluation process. Conversations about implementation difficulties were important, not threatening, to teachers. Discussions about instruction were part of the process, not add-ons. In a supportive local context, teachers have ownership over their respective portfolio implementations, and they take responsibility and pride in them. Teachers are supported and feel supported in their ef-

forts. The process is valued as much as the portfolio product.

The addition of a compatible cross-site (external) perspective strengthens each of the components—portfolio evidence, evaluation process, and professional development. Because, in this case, the external perspective was perceived as supportive, there was no "coaching" for the evaluation sessions or time spent by the teachers or students perfecting portfolio artifacts. There was no fear of results. With an external perspective, teachers have an opportunity to expand their understandings of important learning outcomes, explore new instructional techniques, and anchor their expectations for student performance against common public standards. An external perspective provides a sounding board, a point of comparison, for locally developed outcomes, strategies, and standards. It provides an opportunity to discover inappropriate or unintentional skewing of expectations that sometimes comes from being too isolated or close to one's own work.

Without supportive internal and external conditions, the results of this study certainly would have been different. Had there been less flexibility in portfolio implementation, more requirements for portfolio contents, more time spent on evaluation, or greater familiarity with the evaluation criteria, we might have achieved higher interrater agreement or had less missing evidence. However, we are equally convinced that the benefits to professional development would not have been as great. The portfolio requirements and the evaluation process would have been too far removed from local classrooms to be meaningful and useful for teachers. Similarly, had there been less commonality in literacy philosophy and knowledge-base, teacher support, or portfolio experience across sites, our findings would not have been as positive. The external perspective was beneficial because it was compatible with the emphasis, experience, and needs of each site. Had there been less compatibility, we are certain that all the components would have been more problematic; portfolio evidence would have been more difficult to interpret, the evaluation process would have been more complex, interrater agreement would have been lower, and professional development would have been minimized.

We realize that some will take exception to our model of a classroom-based portfolio assessment system with its emphasis on clearly stated outcomes and some required portfolio evidence (e.g., Carini, 1975; Hansen, 1994). Others will question the amount of support, cost, and time needed to implement such a project. Others will challenge the evaluation process and credibility of results. And others will doubt the need for an external perspective for locally developed portfolios. However, with all the elements in place, we believe the process is a strong one—it has system validity (Frederikson & Collins, 1989; Messick, 1994). This assessment process promotes valued changes in teaching and learning; this is the ultimate goal of assessment.

Taken together, the process and products of this study point to the naiveté of mandating portfolio implementation and expecting immediate results. Certainly, states and school districts can mandate teachers' use of portfolios, but they cannot mandate their effectiveness for assessment nor their positive effects on

teaching and learning. The challenges of our experiences most assuredly will be magnified with mandated portfolio assessment, perhaps beyond the point of finding the process or the results beneficial to any stakeholders—teachers, students, parents, administrators, or policy-makers. Instead, we argue for a portfolio system that includes attention to all the components—portfolio evidence, evaluation process, and professional development—all in supportive local and external contexts. Without such a system, portfolio assessment will be disappointing, ineffective, and doomed to failure. With such a system, it can contribute to high quality education for all children.

Author Note. We are indebted to the teachers and administrators associated with the Bellevue and KEEP portfolio projects for their collaboration, commitment, and support.

References

Applebee, A. N. (1991). Environments for language teaching and learning: Contemporary issues and future directions. In J. Flood, J. M. Jensen, D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 549–556). New York: Macmillan.

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practices, 12*(1), 36–44.

Aschbacher, P. R. (1994). Helping educators to develop and use alternative assessments: Barriers and facilitators. *Educational Policy, 8*, 202–223.

Asam, C., Au, K., Blake, K., Carroll, J., Jacobson, H., & Scheu, J. (1993). *Literacy curriculum guide*. Honolulu, HI: Kamehameha Schools Bernice Pauaki Bishop Estate, Early Education Division.

Au, K. H. (1994). Portfolio assessment: Experiences at the Kamehameha Elementary Education Program. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

Au, K. H., & Asam, C. A. (in press). Improving the achievement of low-income students of diverse backgrounds. In M. Graves, B. Taylor, & P. van den Broek, *The first R: A right of all children*. New York: Teachers College Press.

Au, K. H., Scheu, J. A., Kawakami, A. J., & Herman, P. A. (1990). Assessment and accountability in a whole literacy curriculum. *The Reading Teacher, 43*, 574–578.

Bird, T. (1988). The schoolteacher's portfolio: An essay on possibilities. In J. Millman & L. Darling-Hammond (Eds.), *Handbook of Teacher Evaluation: Elementary and Secondary Personnel* (pp. 241–256).

Calfee, R. C., & Perfumo, P. (Eds.). (1996). *Writing portfolios: Policy and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Calkins, L. M. (1994). *The art of teaching writing* (2nd ed.). Portsmouth, NH: Heinemann.

Carini, P. (1975). *Observation and description: An alternative methodology for the investigation of human phenomena*. Grand Forks, ND: Center for Teaching and Learning, University of North Dakota.

Chittenden, E., & Spicer, W. (1993). *The South Brunswick literacy portfolio project*. Paper presented at the New Standards Project: English/Language Arts Portfolio Meeting, Minneapolis, MN.

Darling-Hammond, L., & Ancess, J. (1993). Authentic assessment and teachers' professional development. *Resources for restructuring: Newsletter of Center for Restructuring Education, Schools, and Teaching*, 1–6.

Frederikson, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Freedman, S. W. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. *Educational Assessment*, 1, 27–52.

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Techincal, Report No. 337). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1993). *"Whose work is it?" A question for the validity of large-scale portfolio assessment* (CSE Techincal Report No. 363). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Graves, D. (1983). *Writing: Teachers and children at work*. Exeter, NH: Heinemann.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), *Expanding student assessment* (pp. 142–163). Alexandria, VA: Association for Supervision and Curriculum Development.

Hansen, J. (1994). Literacy portfolios: Windows on potential. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 26–40). Newark, DE: International Reading Association.

Howard, K. (1990). Making the writing portfolio real. *Quarterly of the National Writing Project and the Center for the Study of Writing and Literacy, 12*(2), 4–7, 27.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*, 509–528.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (Interim Report). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C. (1992). Using student portfolios for a public accounting. *The School Administrator, 49*(11), 8–15.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5–21.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Moss, P. (1994). Can there be validity without reliability. *Educational Researcher, 23*(2), 5–12.

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12–21.

Nystrand, M., Cohen, A. S., & Martinez, M. N. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, 1, 53–70.

Pelavin, S. (1991). *Performance assessments in the states*. Washington, DC: Pelavin Associates.

Roser, N. L., & Martinez, M. G. (Eds.). (1995). *Book talk and beyond: Children and teachers respond to literature*. Newark, DE: International Reading Association.

Short, K. G., & Pierce, K. M. (Eds.). (1990). *Talking about books: Creating literate communities*. Portsmouth, NH: Heinemann.

Tierney, R. J., Carter, M. A., & Desai, L. (1991). *Portfolio assessment in the reading-writing classroom*. Norwood, CA: Christopher-Gordon.

Valencia, S. (1996). *Portfolios in action*. Manuscript in preparation.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *The Reading Teacher, 43*, 338–340.

Valencia, S. W. (1991). Portfolios: Panacea or Pandora's box? In *Educational Performance Testing*. Chicago: Riverside Publishing Company.

Valencia, S. W., & Calfee, R. C. (1991). The development and use of literacy portfolios for students, classes, and teachers. *Applied Measurement in Education, 4*, 333–345.

Valencia, S. W., & Place, N. (1994a). Literacy portfolios for teaching, learning, and accountability: The Bellevue literacy assessment project. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 134–156). Newark, DE: International Reading Association.

Valencia, S. W., & Place, N. (1994b). Portfolios: A process for enhancing teaching and learning. *The Reading Teacher, 47*, 66–69.

Weaver, C. (1990). *Understanding whole language: Principles and practices*. Portsmouth, NH: Heinemann.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*(7).

Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 79*, 703–713.

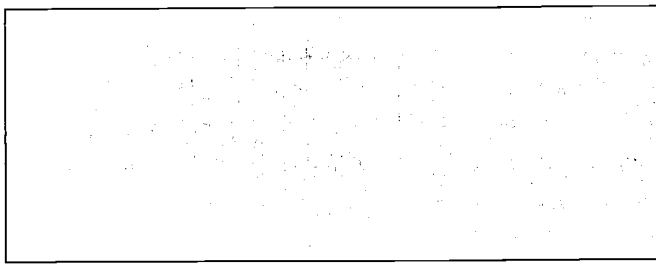
Wiggins, G. (1989b). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41.

Wiggins, G. P. (1993). *Assessing student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wixson, K. K., Valencia, S. W., & Lipson, M. Y. (1994). Issues in literacy assessment: Facing the realities of internal and external assessment. *Journal of Reading Behavior, 26*, 315–337.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(7), 35–39.

Wolf, D. P., LeMahieu, P. G., & Eresh, J. T. (1992). Good measure: Assessment in service to education. *Educational Leadership, 49*(8), 8–13.



NRRC National Reading Research Center

318 Aderhold, University of Georgia, Athens, Georgia 30602-7125
3216 J. M. Patterson Building, University of Maryland, College Park, MD 20742

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").